Two hours

Formula Sheet attached for use with Question 4.

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Text Mining

Date: Friday 1st June 2011
Time: 09:45 - 11:45

Please answer any THREE Questions from the FIVE questions provided

Each question is worth 20 marks

Do not answer more than the required number of questions:
Additional answers will not be marked

Clearly cross out anything you do not wish to be marked.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are
not programmable and do not store text.
1.  

a) Consider the following table, based on a collection where the number of documents \( N = 1000000 \):

<table>
<thead>
<tr>
<th>Term</th>
<th>Document frequency of term (df)</th>
<th>Inverse document frequency of term (idf)</th>
</tr>
</thead>
<tbody>
<tr>
<td>calpurnia</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Animal</td>
<td>100</td>
<td>4</td>
</tr>
<tr>
<td>Sunday</td>
<td>1000</td>
<td>3</td>
</tr>
<tr>
<td>Fly</td>
<td>10000</td>
<td>2</td>
</tr>
<tr>
<td>Under</td>
<td>100000</td>
<td>1</td>
</tr>
<tr>
<td>The</td>
<td>1000000</td>
<td>0</td>
</tr>
</tbody>
</table>

i) What is the reasoning behind the calculation of inverse document frequency (idf)?

(1 mark)

ii) I now calculate how many times each term appears in each document, to yield \( tf_{t,d} \). I then calculate a tf-idf weight for each term in each document. I note that, for some term, its weight is less for document \( A \) than for document \( B \). What is this telling me? How would I use this information to rank documents for relevance in response to a query?

(3 marks)

b) Explain two ways in which cosine similarity scores are used with the vector space model.

(2 marks)

c) I form a new document by taking a copy of a document and appending it to itself. This new document evidently has the same semantic content as the original. What steps do I have to take to ensure that a cosine similarity calculation does not cause me to think that these two documents are quite different?

(1 mark)

d) The formula for Precision is:

\[
\frac{TP}{(TP+FP)}
\]

and that for Recall is:

\[
\frac{TP}{(TP+FN)}
\]

where ‘\( T \)’ is ‘true’, ‘\( F \)’ is ‘false’, ‘\( P \)’ is ‘positives’ and ‘\( N \)’ is ‘negatives’.

[Question 1 continues on the following page]
Using the above formulae, give the Precision and Recall of a system that returns 8 relevant documents and 10 non-relevant documents for a collection where there are known to be 20 relevant documents. Show your working.

(2 marks)

e) What is the advantage of using F measure, the harmonic mean of Precision and Recall, rather than the arithmetic mean?

(2 marks)

f) Classify the following examples according to the type of ambiguity they display

i) John swerved and hit a tree riding his bike.

ii) The old man the boat.

iii) I saw her duck.

iv) I saw the man in the park under the tree with the telescope

v) Coming round the corner, a magnificent lake came into view.

vi) The soldiers shot the women and they fell down.

(3 marks)

g) You are asked to make recommendations on using either a morphological analyser or a stemmer in the following frameworks:

- An information retrieval system
- An information extraction system

Discuss advantages and disadvantages of each within these frameworks, and give your recommendations, with justifications.

(3 marks)

h) Syntactic analysis of the sentences Jim killed the goat from Hannah and Jim stole the goat from Hannah does not yield the same number of phrase structure trees for each sentence. Explain why this is the case, what linguistic notion is involved, and how this notion can be used in natural language processing, giving examples.

(3 marks)
2.

a) For a sequence of observations, \( O = \{o_1, o_2, o_3 \ldots o_n\} \) and class of labels, \( L = \{l_1, l_2, l_3, \ldots l_n\} \), describe the most probable label sequence, using a Hidden Markov Model (HMM).

(2 marks)

b) Write down the Viterbi equation for the best sequence of labels \( L \) for observations \( O \), as described in 2a) above.

(1 mark)

c) i) What are the assumptions made when using HMMs?

(1 mark)

ii) What limitations are due to these assumptions and how can one overcome them?

(2 marks)

d) i) Describe Conditional Random Fields (CRFs).

(1 mark)

ii) How do they differ from HMMs?

(1 mark)

e) Discuss briefly aspects of training a model in the CRF framework.

(2 marks)

f) Briefly explain, with example data, the role of each of the following types of analysis component in a text mining pipeline:

i) chunker

ii) gazetteer

iii) PDF to text converter

iv) part of speech tagger

v) sentence splitter

vi) tokenizer

(2 marks)

g) State the sequence in which you would expect to apply the components of 2f), explaining your choice, and noting any inherent ordering constraints that may apply.

(2 marks)

[Question 2 continues on the following page]
[Question 2 continues from the previous page]

h) Explain the BIO and BILOU notations and how they are used in representing text mining annotations.

(2 marks)

i) What are the relative merits of in-line and standoff annotations in representing the output of text mining components?

(2 marks)

j) Explain the Transformation-Based Learning algorithm for training sequence taggers, noting the role of the corpus, the lexicon, the initial and intermediate states, the rule templates and the guesser.

(2 marks)
3.

a)  

i) Briefly explain the difference between chunking and (full) parsing.  

(2 marks)

ii) Suggest an application where chunking would be the more appropriate technique.  

(1 mark)

b)  

Consider the following grammar and lexicon:

\[
S \rightarrow NP \ VP  \\
NP \rightarrow N  \\
NP \rightarrow N \ N  \\
VP \rightarrow V \ NP  \\
V \rightarrow \text{cuts}  \\
V \rightarrow \text{surprise}  \\
N \rightarrow \text{government}  \\
N \rightarrow \text{party}  \\
N \rightarrow \text{cuts}  \\
N \rightarrow \text{surprise} 
\]

i) Show, by constructing two parse trees (or labelled bracketings) that the string "government cuts surprise party" is ambiguous according to the above grammar.  

(2 marks)

ii) By reference to the same grammar and lexicon, an Earley parser has read some lexical entries, and its chart is in the state shown below. Reproduce this chart, showing six additional edges that can be created by triggering rules or extending active edges.  

(3 marks)
c) Explain how a treebank parser is constructed, and how its operation and outputs differ from those of a nondeterministic parser applying a broad coverage context-free grammar created by a trained linguist.

(2 marks)

d) Compare and contrast a phrase structure tree and a dependency tree, discussing why the latter is often preferred for applications in text mining.

(2 marks)

e) i) What motivates the use of a probabilistic context-free grammar (PCFG)?

(2 marks)

ii) Briefly explain how one is constructed from a treebank.

(2 marks)

f) What syntactic knowledge can be used in resolving the ambiguity in word senses or named entities? In your answer, give at least one example explaining how the syntactic knowledge you identify can help and also explain the limitations of such knowledge for this purpose.

(2 marks)

g) Between the words eat and find, state which you would expect to be more effective in selectional restriction-based sense disambiguation. Justify your answer.

(2 marks)
4.  

a)  *For this question part, consult the provided formula sheet.*

A corpus consists of 20000 bigrams, where *new* occurs in 400 bigrams and *York* occurs in 50 bigrams. The bigram *new York* occurs 25 times.

i)  Compute the pointwise mutual information of the bigram *new York*.  

(2 marks)

ii)  Decide if the co-occurrence of *new* and *York* is random or not using the *t test*. Show your working. (The critical value for a confidence level $\alpha=0.005$ is 2.576.)

(3 marks)

iii)  Compute the observed values contingency table of the $X^2$ (chi-square) test.

(1 mark)

iv)  Compute the expected values contingency table of the $X^2$ (chi-square) test.

(2 marks)

v)  Decide if the co-occurrence of “*new*” and “*York*” is random or not using the $X^2$ (chi-square) test. Show your working. (For 1 degree of freedom and at a probability level of $\alpha=0.05$ the critical value is 3.841.)

(3 marks)


(2 marks)

c)  Discuss the extent to which Okazaki’s approach offers benefits over letter matching approaches.

(2 marks)

d)  Compare and contrast mutual information with the C-Value measure, discussing to what extent these measures are useful in term recognition.

(2 marks)

e)  Briefly describe how the naive Bayes classification algorithm can be adopted to the task of recognising named entities in text, assuming a gold-standard corpus is available, where occurrences of entities of interest have been manually tagged.

(3 marks)
5.

a) Explain the role of the type system in the Apache Unstructured Information Management Architecture (UIMA).

(3 marks)

b) Many component libraries for text mining developed independently of UIMA have been "wrapped" for use in the UIMA framework. What does a developer do in order to achieve this?

(3 marks)

c) You have been appointed to lead a team that will develop text mining software applications. Your managers want to know what standards and frameworks you are adopting. Write a memo to them explaining why you have chosen to adopt either UIMA or U-Compare (or any plausible alternative to these).

(4 marks)

d) Consider the following quotations:

“Text mining will catch and eventually dwarf traditional information retrieval.” (Rao, Queue, 2004)

“Web and enterprise users will become accustomed to interacting with and exploring information, and there will be no going back to plain-old keyword search and low-value hit lists of search results.” (Grimes, Alta Plana)

“Text mining is frequently viewed as an information technology tool – build it, and people will use it.” (Lavengood & Kiser, Online, 2011)

Taking these as a starting point, discuss the future of text mining. Justify your views and conclusions, giving appropriate examples to back up your arguments.

(10 marks)
Pointwise Mutual Information (PMI)

\[ I(x, y) = \log_2 \frac{P(x, y)}{P(x) P(y)} = \log_2 \frac{P(x|y)}{P(x)} = \log_2 \frac{P(y|x)}{P(y)} \]  

where \( x \) and \( y \) are events.

**T statistic**

\[ t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \]  

where \( \bar{x} \) is the sample mean, \( s^2 \) the sample variance, \( N \) the sample size, and \( \mu \) the mean of the distribution.

<table>
<thead>
<tr>
<th>( \nu ) ( \backslash \alpha )</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
<th>0.0025</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
<td>127.3</td>
<td>318.3</td>
<td>636.6</td>
</tr>
<tr>
<td>2</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.25</td>
<td>14.09</td>
<td>22.33</td>
<td>31.60</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
<td>7.453</td>
<td>10.21</td>
<td>12.92</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>5.598</td>
<td>7.173</td>
<td>8.610</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>4.773</td>
<td>5.893</td>
<td>6.869</td>
</tr>
<tr>
<td>6</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>4.317</td>
<td>5.208</td>
<td>5.959</td>
</tr>
<tr>
<td>7</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>4.029</td>
<td>4.785</td>
<td>5.408</td>
</tr>
<tr>
<td>8</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
<td>3.833</td>
<td>4.501</td>
<td>5.041</td>
</tr>
<tr>
<td>9</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
<td>3.690</td>
<td>4.297</td>
<td>4.781</td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>3.581</td>
<td>4.144</td>
<td>4.587</td>
</tr>
<tr>
<td>20</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.153</td>
<td>3.552</td>
<td>3.850</td>
</tr>
<tr>
<td>30</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.030</td>
<td>3.385</td>
<td>3.646</td>
</tr>
<tr>
<td>40</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>2.971</td>
<td>3.307</td>
<td>3.551</td>
</tr>
<tr>
<td>50</td>
<td>1.676</td>
<td>2.009</td>
<td>2.403</td>
<td>2.678</td>
<td>2.937</td>
<td>3.261</td>
<td>3.496</td>
</tr>
<tr>
<td>60</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>2.915</td>
<td>3.232</td>
<td>3.460</td>
</tr>
<tr>
<td>80</td>
<td>1.664</td>
<td>1.990</td>
<td>2.374</td>
<td>2.639</td>
<td>2.887</td>
<td>3.195</td>
<td>3.416</td>
</tr>
<tr>
<td>100</td>
<td>1.660</td>
<td>1.984</td>
<td>2.364</td>
<td>2.626</td>
<td>2.871</td>
<td>3.174</td>
<td>3.390</td>
</tr>
<tr>
<td>120</td>
<td>1.658</td>
<td>1.980</td>
<td>2.358</td>
<td>2.617</td>
<td>2.860</td>
<td>3.160</td>
<td>3.373</td>
</tr>
<tr>
<td>( \infty )</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>2.807</td>
<td>3.090</td>
<td>3.291</td>
</tr>
</tbody>
</table>

Table 1: \( t \) distribution table (\( \nu \): degrees of freedom)
$X^2$ statistic

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

(3)

where $O$ denotes the contingency table of observed values and $E$ the contingency table of expected values.

<table>
<thead>
<tr>
<th>$\nu \setminus \alpha$</th>
<th>0.250</th>
<th>0.100</th>
<th>0.050</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.3233</td>
<td>2.7055</td>
<td>3.8415</td>
<td>5.0238</td>
<td>6.6349</td>
<td>7.8794</td>
</tr>
<tr>
<td>2</td>
<td>2.7726</td>
<td>4.6052</td>
<td>7.3776</td>
<td>9.2103</td>
<td>10.5966</td>
<td>10.9646</td>
</tr>
<tr>
<td>4</td>
<td>5.3853</td>
<td>7.7944</td>
<td>11.1432</td>
<td>13.2767</td>
<td>15.0862</td>
<td>16.7496</td>
</tr>
<tr>
<td>7</td>
<td>9.0371</td>
<td>12.0170</td>
<td>16.0128</td>
<td>18.4753</td>
<td>20.2774</td>
<td>22.1505</td>
</tr>
<tr>
<td>10</td>
<td>12.5485</td>
<td>15.9871</td>
<td>18.3070</td>
<td>20.4830</td>
<td>23.0925</td>
<td>25.1829</td>
</tr>
<tr>
<td>14</td>
<td>17.1160</td>
<td>21.0414</td>
<td>23.6879</td>
<td>26.1189</td>
<td>29.1412</td>
<td>31.3193</td>
</tr>
<tr>
<td>16</td>
<td>19.3685</td>
<td>23.5183</td>
<td>26.2923</td>
<td>28.8435</td>
<td>31.9999</td>
<td>34.2671</td>
</tr>
<tr>
<td>18</td>
<td>21.6049</td>
<td>25.9842</td>
<td>28.8930</td>
<td>31.5263</td>
<td>34.8053</td>
<td>37.1564</td>
</tr>
<tr>
<td>20</td>
<td>23.8279</td>
<td>28.4198</td>
<td>31.4043</td>
<td>34.1601</td>
<td>37.6626</td>
<td>39.9980</td>
</tr>
<tr>
<td>22</td>
<td>26.0392</td>
<td>30.8132</td>
<td>33.9243</td>
<td>36.7071</td>
<td>40.2893</td>
<td>42.7965</td>
</tr>
<tr>
<td>24</td>
<td>28.2415</td>
<td>33.1924</td>
<td>36.4150</td>
<td>39.3640</td>
<td>42.9798</td>
<td>45.5581</td>
</tr>
<tr>
<td>26</td>
<td>30.4345</td>
<td>35.5631</td>
<td>38.8514</td>
<td>41.9231</td>
<td>45.6148</td>
<td>48.2898</td>
</tr>
<tr>
<td>28</td>
<td>32.6249</td>
<td>37.9159</td>
<td>41.3714</td>
<td>44.6074</td>
<td>48.2784</td>
<td>50.9338</td>
</tr>
<tr>
<td>30</td>
<td>34.7997</td>
<td>40.2560</td>
<td>43.7729</td>
<td>46.9792</td>
<td>50.8921</td>
<td>53.6719</td>
</tr>
</tbody>
</table>

Table 2: $\chi^2$ distribution table ($\nu$: degrees of freedom)